A Stitch in Time Saves Nine: A Train-Time Regularizing Loss for Improved Neural Network Calibration

Ramya Hebbalaguppe[1,2,*] Jatin Prakash[1,*] Neelabh Madan[1,*] Chetan Arora[1]

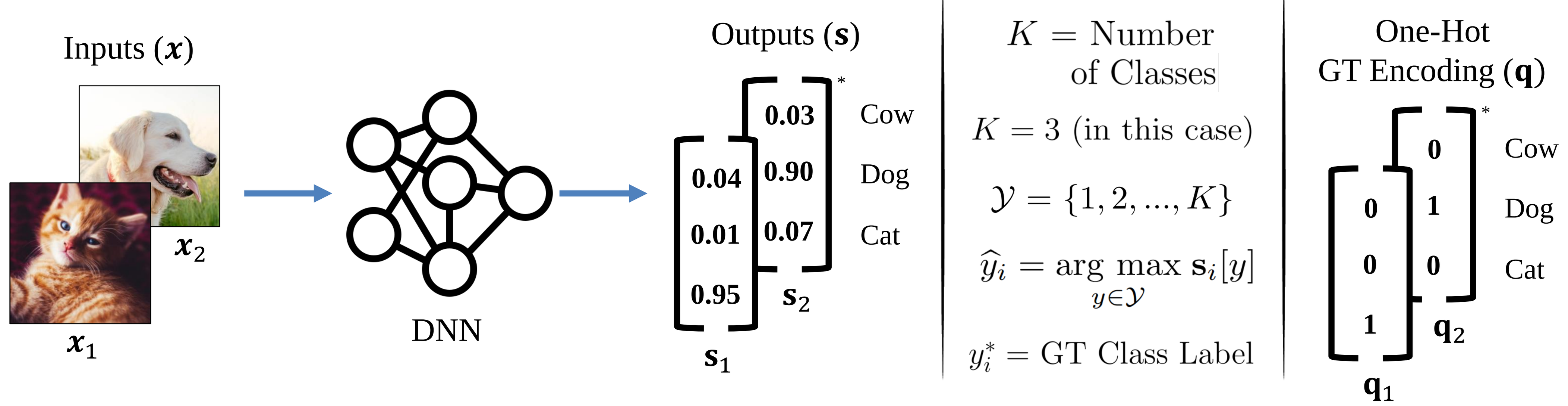
[1]Indian Institute of Technology Delhi, India [2]TCS Research, India

Highlights

- [Novelty] We propose an auxiliary loss to overcome miscalibration
- [Multi-class Calibration] Entire probability vector (all K classes) taken into account
- [Powerful Regularizer] Models trained using our method are relatively well calibrated even under domain/dataset drift
- [Superior Calibration] Outperforms SOTA methods on various datasets and models
- [Beyond Image Classification] Promising results in semantic segmentation in images and NL classification tasks

Understanding Calibration

"If a calibrated model predicts an event with 0.7 confidence, then 70% of the times the event transpires"



Top-Label Calibration

$$\mathbb{P}(\widehat{y}_i = y_i^* \mid \mathbf{s}[\widehat{y}_i] = p) = p$$

Multi-class Calibration

$$\mathbb{P}(y = y_i^* \mid \mathbf{s}_i[y] = p) = p \quad \forall\, y \in \mathcal{Y}$$

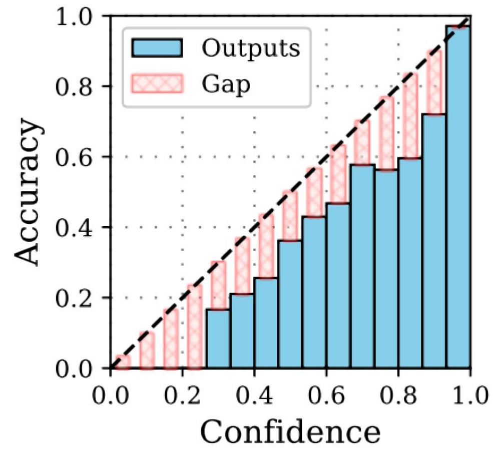
Problem: Modern Neural Networks are neither top-label nor multi-class calibrated

Measuring Calibration

1. Quantitative Measures

- [ECE] Expected Calibration Error: It calculates the absolute difference between the model's accuracy and confidence. It captures the information about top-label calibration.
- [SCE] Static Calibration Error: A simple class-wise extension to ECE that captures multi-class calibration

2. Reliability Diagrams



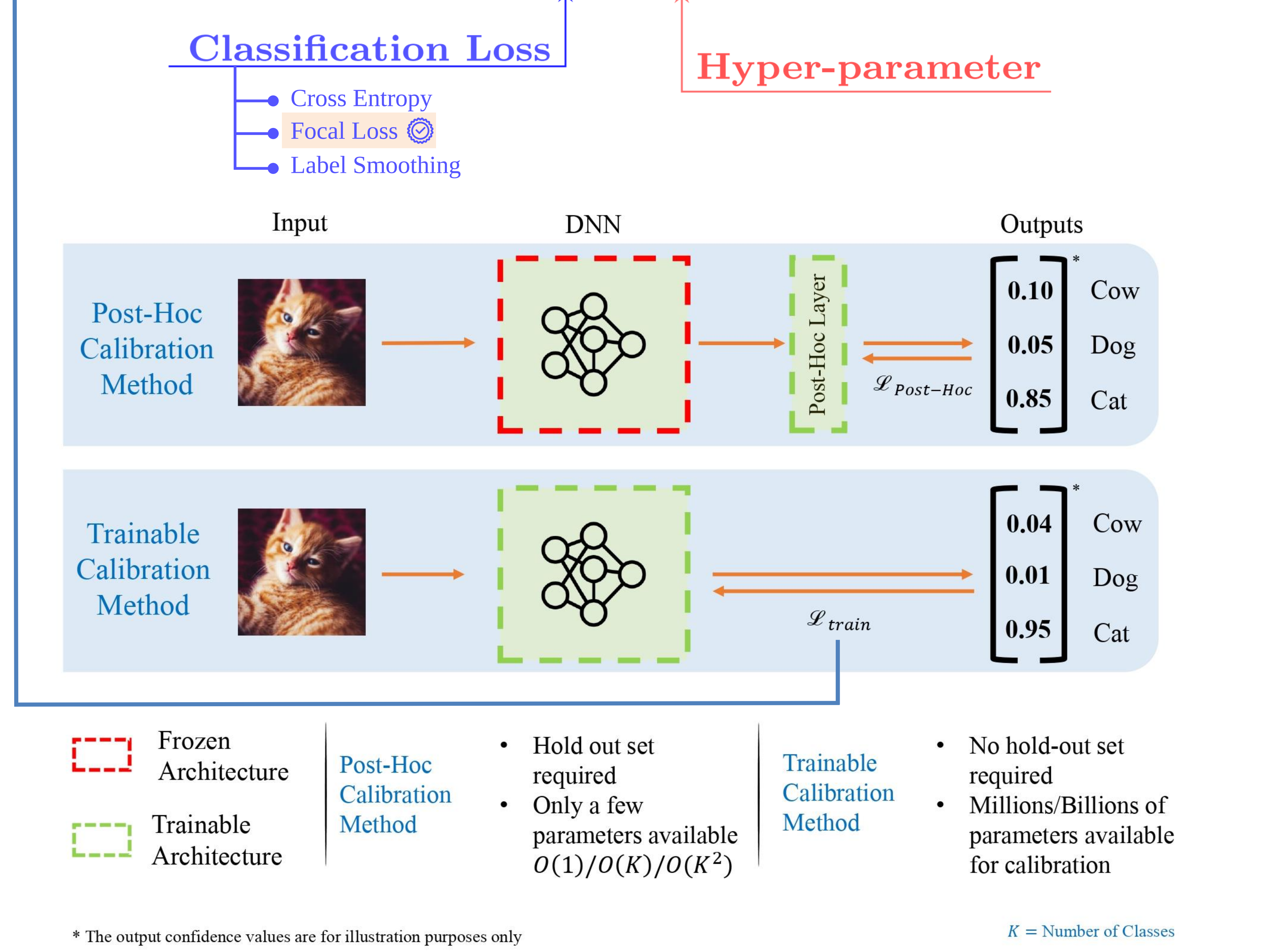
Proposed Solution

We propose a novel train-time regularizing auxiliary loss function called Multi-class Difference in Confidence and Accuracy (MDCA)

$$\mathcal{L}_{\text{MDCA}} = \frac{1}{\underbrace{K}_{\text{Number of classes}}} \sum_{j=1}^{K} \left| \overbrace{\frac{1}{\underbrace{N_b}_{\text{Number of samples in a mini-batch}}} \sum_{i=1}^{N_b} \mathbf{s}_i[j]}^{\text{Avg. Confidence}} - \overbrace{\frac{1}{N_b} \sum_{i=1}^{N_b} q_i[j]}^{\text{Avg. Count}} \right|$$

Our Proposed Auxiliary Loss

$$\mathcal{L}_{\text{total}} = \underbrace{\mathcal{L}_C}_{\text{Classification Loss}} + \underbrace{\beta}_{\text{Hyper-parameter}} \cdot \mathcal{L}_{\text{MDCA}}$$



Paper and Code:
https://github.com/mdca-loss

Experimental Results

1. Superior performance against trainable calibration methods

Dataset	Model	BS [2] SCE	BS [2] ECE	BS [2] TE	DCA [31] SCE	DCA [31] ECE	DCA [31] TE	MMCE [26] SCE	MMCE [26] ECE	MMCE [26] TE	FLSD [37] SCE	FLSD [37] ECE	FLSD [37] TE	Ours (FL+MDCA) SCE	Ours (FL+MDCA) ECE	Ours (FL+MDCA) TE
CIFAR10	ResNet32	6.60	2.92	7.76	8.41	4.00	**7.06**	8.17	3.31	8.41	9.48	4.41	7.87	**3.22**	**0.93**	7.18
	ResNet56	5.44	2.17	7.75	7.59	3.38	**6.53**	9.11	3.71	8.23	7.71	3.49	7.04	**2.93**	**0.70**	7.08
CIFAR100	ResNet32	1.97	5.32	33.53	2.82	11.31	29.67	2.79	11.09	31.62	1.77	1.69	32.15	**1.72**	**1.49**	**31.58**
	ResNet56	1.86	4.69	30.72	2.77	9.29	43.43	2.35	8.61	**28.75**	1.71	1.90	29.11	**1.60**	**0.72**	29.8
SVHN	ResNet20	2.12	**0.45**	**3.56**	4.29	2.02	3.83	9.18	4.34	4.12	18.98	9.37	4.10	**1.90**	0.47	3.92
	ResNet56	2.18	0.66	**3.25**	2.16	0.49	3.32	9.69	4.48	4.26	26.15	13.23	3.65	**1.51**	**0.23**	3.85
Mendeley V2	ResNet50	117.6	3.75	18.43	145.1	8.29	17.47	130.4	**3.45**	**15.06**	104.3	9.64	19.71	**85.68**	4.81	17.95
Tiny-ImageNet	ResNet34	1.53	7.79	43.00	2.11	17.40	**36.68**	1.62	9.71	40.75	1.18	**1.91**	37.01	**1.17**	1.99	37.49
20 Newsgroups	Global-Pool CNN	725.82	13.71	**25.93**	719.83	15.30	28.07	731.31	12.69	28.63	940.70	**4.52**	30.80	**487.82**	16.55	27.88

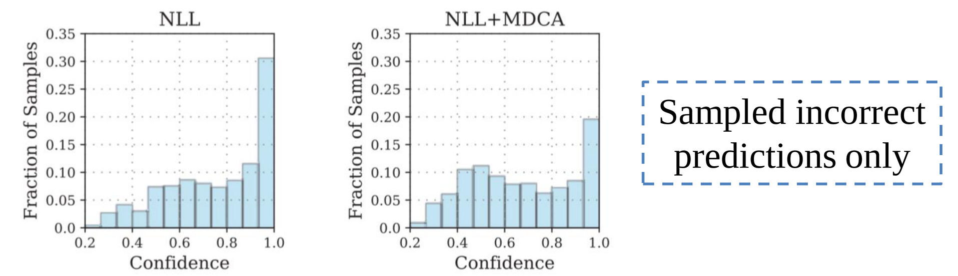
2. Superior class-wise calibration

Method	Classes 0	1	2	3	4	5	6	7	8	9
Cross Entropy	**0.20**	0.62	0.33	0.65	0.23	0.36	0.25	0.26	**0.21**	0.41
Focal Loss [32]	0.30	0.48	0.41	**0.18**	0.38	0.19	0.33	0.36	0.32	0.30
LS [38]	1.63	2.60	2.54	1.90	1.91	1.74	1.73	1.75	1.63	1.58
Brier Score [2]	0.23	0.28	0.40	0.45	0.25	0.26	0.25	0.27	0.21	0.37
MMCE [26]	1.78	2.35	2.12	2.00	1.74	1.87	1.65	1.76	1.70	1.84
DCA [31]	0.31	0.70	0.40	0.72	0.31	0.46	0.35	0.35	0.37	0.36
FLSD [37]	1.52	3.24	2.74	2.15	1.79	1.82	1.84	1.62	1.54	1.38
Ours (FL+MDCA)	0.22	**0.16**	**0.24**	0.25	**0.22**	**0.16**	**0.16**	**0.17**	0.25	**0.20**

3. Performance under dataset drift

Method	Art	Cartoon	Sketch	Average
NLL	6.33	17.95	15.01	13.10
LS [38]	7.80	11.95	**10.88**	10.21
FL [32]	8.61	16.62	10.94	12.06
Brier Score [2]	6.55	13.19	15.63	11.79
MMCE [26]	6.35	15.70	17.16	13.07
DCA [31]	7.49	18.01	14.99	13.49
FLSD [37]	8.35	13.39	13.86	11.87
Ours (FL+MDCA)	**6.21**	**11.91**	11.08	**9.73**

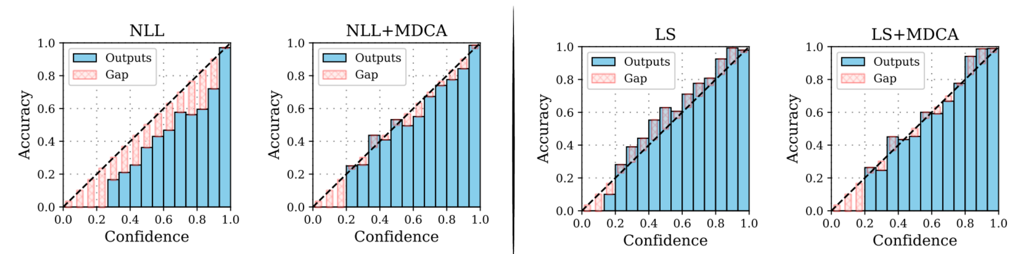
4. Mitigating overconfident mistakes



5. Performance under data imbalance

Method	CIFAR10 IF-10	CIFAR10 IF-50	CIFAR10 IF-100	SVHN IF-2.7
NLL	18.44	32.21	31.04	3.43
FL [32]	14.65	29.67	28.89	2.54
LS [38]	14.88	26.30	**20.79**	18.80
BS [2]	15.74	33.57	29.01	2.12
MMCE [26]	15.10	29.05	21.56	9.18
FLSD [37]	16.05	31.35	30.28	18.98
DCA [31]	18.57	32.81	35.53	4.29
Ours (FL+MDCA)	**11.83**	**22.97**	26.89	**1.90**

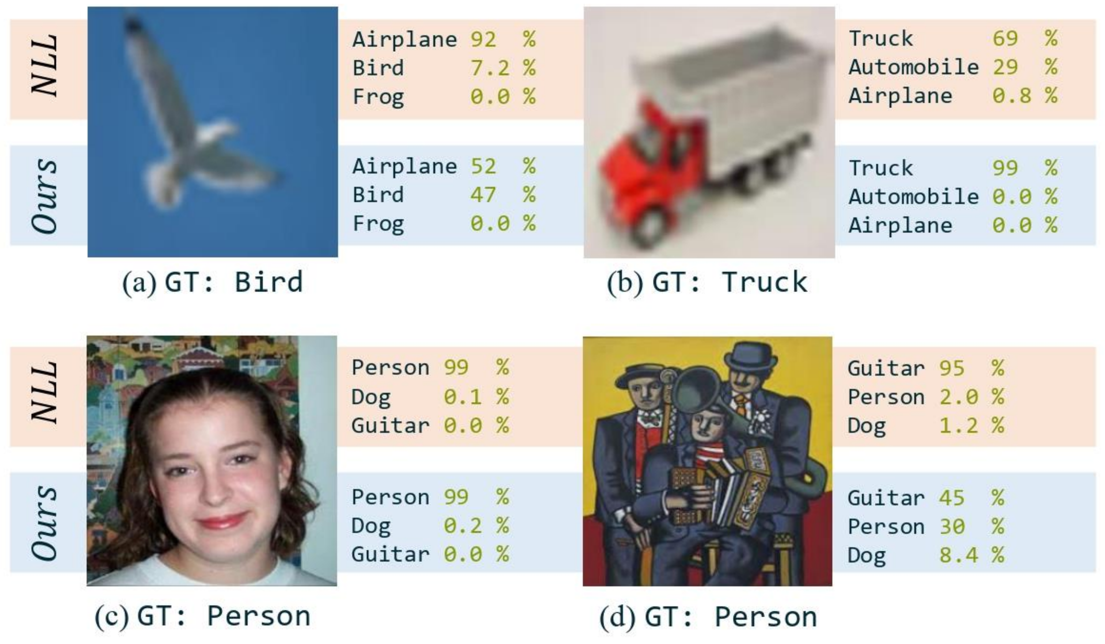
6. Mitigating over/under confidence



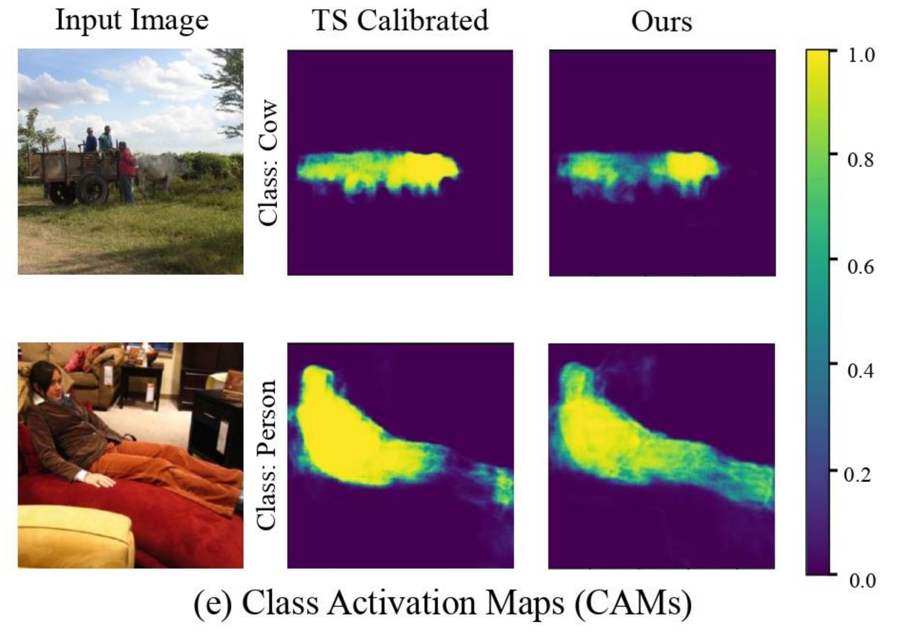
Other results include

- Superior semantic segmentation results
- Superior performance against post-hoc calibration methods

Qualitative Results



(a) GT: Bird (b) GT: Truck (c) GT: Person (d) GT: Person



(e) Class Activation Maps (CAMs)